DOCUMENT RESUME

ED 214 895 SP 019 834

AUTHOR Sirotnik, Kenneth A.
TITLE An Inter-Observer Reliability Study of the SRI Observation System as Modified for Use in a Study of Schooling. A Study of Schooling in the United States. Technical Report Series, No. 27
INSTITUTION California Univ., Los Angeles. Graduate School of Education
SPONS AGENCY Institute for Development of Educational Activities, Dayton, Ohio.
PUB DATE 81
NOTE 39p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Classroom Environment; *Classroom Observation Techniques; Elementary Secondary Education; Experimenter Characteristics; *Observation; *Rating Scales; Student Behavior; *Student Teacher Relationship; Teacher Behavior; *Test Reliability
IDENTIFIERS *SRI Observational System

ABSTRACT

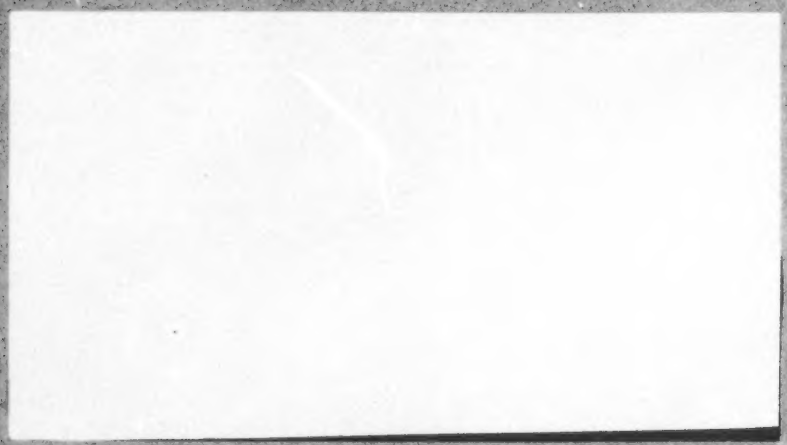
This study examined the objectivity, trainability, and reliability of the SRI observational system (developed at the Stanford Research Institute) as it was modified for use in the "A Study of Schooling" research project. Four instruments in the observation system were tested. The physical environment inventory recorded the architectural arrangement of the classroom, seating and grouping patterns, and materials. The daily summary provided an overview of available space and materials as well as of student and teacher decision making processes. The classroom snapshot provided information about what each person in the classroom was doing at a given point in time and the nature of classroom activities. The five minute interaction was a continuous accounting of how time was spent in the classroom, focusing on the interactive process between teacher and students. By placing two observers in 357 elementary and secondary school classrooms, inter-observer reliability was analyzed. The results indicated that observers could be trained to use the SRI Observational System and that the instruments provided reliable data for analysis of classroom observations. Unreliability occurred mostly for behaviors which were rarely observed. Specific similarities and problems found in the use of each instrument are discussed. Appendices present data on correlations between observers. (CJ)

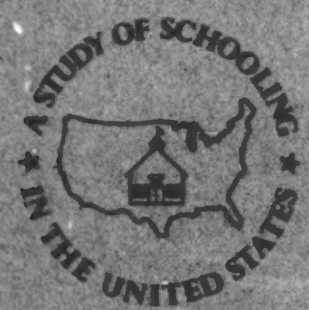
****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED214895

A STUDY OF SCHOOLING

Technical Report Series





"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. GOODLAD

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Distributed by the Laboratory in School and Community Education

Graduate School of Education
University of California
Los Angeles, California 90024

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

SUPPORTED BY

The Danforth Foundation
International Paper Company Foundation
The JDR 3rd Fund
Martha Holden Jennings Foundation
Charles F. Kettering Foundation
Lilly Endowment, Inc.
Charles Stewart Mott Foundation
National Institute of Education
The Needmor Fund
Pedamorphosis, Inc.
The Rockefeller Foundation
The Spencer Foundation
United States Office of Education

Conducted under the auspices of the Institute for Development of Educational Activities (/I/D/E/A/)
A Division of the Charles F. Kettering Foundation

John I. Goodlad, Principal Investigator

AN INTER-OBSERVER RELIABILITY STUDY OF THE SRI OBSERVATION SYSTEM AS MODIFIED FOR USE IN A STUDY OF SCHOOLING

Kenneth A. Sirotnik

Technical Report No. 27

1981

A Study of Schooling is based upon the assumption that improving schools requires knowing what is happening in and around them. A comprehensive data-base of contextual information was obtained from students, teachers, administrators, parents and observers at all grade levels in thirty-eight elementary and secondary purposively sampled schools. *It is strongly recommended that readers of any technical report in this series first read Technical Report No. 1 which outlines the details, scope and limitations of the Study as a whole.*

It must be understood that this series of technical reports does not constitute the Study. Some reports are highly specific "molecular" inquiries while others take a more "molar" view across data sources, schooling levels, etc. Some reports are more methodological in nature arising out of issues in data analysis. Many of the reports quite naturally overlap in data analysed and interpretations rendered. Some authors have approached their task as consisting mostly of data description with little discussion beyond the presentation of the data. Others have ventured further into the realm of interpretation and speculation. *It must be further understood that data-based inferences can and do differ among researchers who come at the data from differing points-of-view.* Authors, therefore, are duly acknowledged for each report and are responsible for the material presented therein.

An Inter-Observer Reliability Study
of the SRI Observation System as Modified for Use
in A Study of Schooling

Introduction

Fortunately, the educational research community has become increasingly aware of the fact that more than one "reliability coefficient" may exist, each based upon differing circumstances under which dependable assessments are required. In the case of observational assessment, the changing circumstances due to different observers and/or different observational occasions are particularly important. Given the impetus of early work adapting the principles of analysis of variance for use in psychometric analysis (e.g., Lindquist, 1953; Medley and Mitzel, 1963; and Cronbach, Rajaratnum and Gleser, 1963), the methods for obtaining one or more relevant reliability indices were synthesized under the rubric of generalizability theory by Cronbach, Gleser, Nanda and Rajaratnum (1972).

As applied to the observation of classroom processes, a simple design would involve at least three facets (or factors): the classes observed, the observers, and the occasions of observation (e.g., days, periods, etc.). Three reliability coefficients emerge with potential utility: (1) the correlation between different observers on the same occasion--inter-observer agreement, (2) the correlation between observations on different occasions by the same observer--stability for the single observer, and (3) the correlation between observations on different occasions for different observers--stability across different observers.

It has been duly noted (Medley and Mitzel, 1963 and Shavelson and Demsey, 1976) that the third coefficient is the most useful index of reliability from a practical standpoint. Ordinarily, the variance component

due to occasions is substantially greater than that due to observers. Moreover, the validity of observational data as predictors of educational outcomes can be expected to increase as the observations increase in number and representativeness. In practice, the same observer will not always observe the same class on every occasion. These practical concerns lead to stability across different observers as ultimately the most useful reliability index.

Nevertheless, the other reliability forms can also be useful. Coefficients of stability holding observer constant are useful reliability approximations, especially when interrater (or inter-observer) reliability is high. The inter-observer reliability is particularly useful as an index of the objectivity of training. Certainly, observation systems with even low to moderate inter-observer reliability can be sufficiently valid predictors so long as observations are made over enough occasions (a typically difficult and/or expensive effort). However, high inter-observer reliability is certainly a desirable ingredient if possible and a necessary index for evaluating the consistency of training observers.

It should be noted here that other types of "reliability" coefficients can be computed which do not readily fit into the generalizability framework. For example, "criterion-referenced" agreement coefficients between the observers' ratings and the "correct" ratings, as judged by "experts," can be constructed. (See, for example, the confusability matrix approach developed by Baker, Giesen and Norwood, 1974, and the application of this approach to the SRI observation system by Stallings and Giesen, 1974.) Again, this is an approach with limited value for describing the reliability of observational data as used in an applied setting requiring predictive validity. Rather, it is an approach yielding important information about

the observational system as being a sufficiently objective and trainable instrument.

This paper is concerned with the objectivity and trainability, as reflected in inter-observer reliability coefficients, of the SRI observational system as modified for use in A Study of Schooling.[1] During the training process, data were collected and analysed to test the accuracy of observers' ratings compared to precoded video-taped and written classroom situations (see Giesen and Sirotnik, 1979). A more realistic test of the observational system, however, allows the inclusion of observer variance in uncontrolled classroom settings. These results will be presented here following a brief description of the observation instrument and the variables to be used in the analyses.

The Observation System and Variable Definition

A more complete description of the instrument and information yielded can be found in the technical report by Giesen and Sirotnik (1979). The observation system is quite complex, and the reader interested in a more in-depth understanding of the instrument than can be presented here is strongly urged to consult this technical report. Briefly, the observation system is a modified version of that developed at the Stanford Research Institute by Jane Stallings and her associates for the evaluation of Project Follow Through (Stallings and Kaskowitz, 1974). Modifications for A Study of Schooling occurred in mainly three ways: (1) it was generalized for use at both elementary and secondary schooling levels, (2) variables were separated out by course content and (3) variables were separated out by classroom contexts: instruction, behavior, routine and the remainder (which was labeled "social").

This study will focus on the data obtained from the four major sections of the instrument--physical environment inventory (PEI), daily summary (DS), classroom snapshot (CS), and five minute interaction (FMI)--with particular emphasis on the latter two sections. The PEI is designed to record the architectural arrangement of the classroom, seating and grouping patterns, furnishings, and materials and equipment. The DS provides an overview of the space and materials available as well as the decision-making processes in evidence by students and teacher.

The CS and FMI sections are considerably more complicated. They occur as pairs four times in a given observation booklet and are recorded in four equal time intervals per day (at the elementary level) or per period (at the secondary level). The classroom snapshot provides information about what each adult (usually a teacher) and student in the classroom is doing, the size of student groups (if any) and the nature of the activities in progress. The typical CS coding task is to "bubble-in" (on an optical scanned format) the following matrix for each relevant activity:

	One Student	Small Groups	Medium Groups	Large Groups	Total Class
T	①②③④⑤	①②③④	①②③	①②	①
A	①②③④⑤	①②③④	①②③	①②	①
C	①②③④⑤	①②③④	①②③	①②	①
I	①②③④⑤	①②③④	①②③	①②	①

The T, A, C and I rows denote "director-type" modalities representing teacher, aide, students cooperating, or students working independently. The column headings denote group sizes (small = 2-6 students; medium = 7-13 students; large = over 13 students) and include individual students and the total class.

When these matrices are crossed with activity types, the three-fold classification of activity-by-director-by-group can describe the whereabouts of every person in the classroom at any point in time (hence the term "snapshot"). A common classroom situation finds the teacher lecturing to the total class and it is recorded as follows:

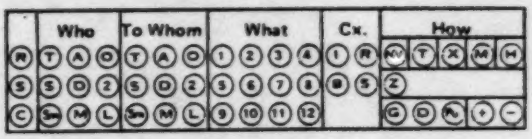
ACTIVITIES		One Student	Small Groups	Medium Groups	Large Groups	Total Class
○ Eng ○ Math ○ Sci ○ S.S. ○ Arts ○ F.L. ○ P.E. ○ Story Time ○ Materials and Equipment	T	①②③④⑤	①②③④	①②③	①②	●
	A	①②③④⑤	①②③④	①②③	①②	①
	S	①②③④⑤	①②③④	①②③	①②	①
2. Explain, Lecture, or Read Aloud						

(The content bubbles enable the observers to record what subject(s) are in progress at the elementary level.) A more complex pattern would require more activity rows for recording. For example, the following CS record indicates that the teacher is demonstrating something to a small group of students, two other small groups are engaged in separate discussions, and the rest of the students in the class are working independently on written assignments (except for one student who is being helped by an aide):

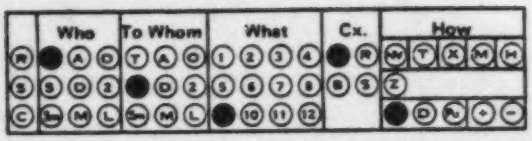
ACTIVITIES		One Student	Small Groups	Medium Groups	Large Groups	Total Class
○ Eng ○ Math ○ Sci ○ S.S. ○ Arts ○ F.L. ○ P.E.	T	①②③④⑤	●②③④	①②③	①②	①
	A	①②③④⑤	①②③④	①②③	①②	①
	S	①②③④⑤	①②③④	①②③	①②	①
3. Demonstration						
○ Eng ○ Math ○ Sci ○ S.S. ○ Arts ○ F.L. ○ P.E.	T	①②③④⑤	①②③④	①②③	①②	①
	A	①②③④⑤	①②③④	①②③	①②	①
	C	①②③④⑤	①●③④	①②③	①②	①
4. Discussion						
○ Eng ○ Math ○ Sci ○ S.S. ○ Arts ○ F.L. ○ P.E.	T	①②③④⑤	①②③④	①②③	①②	①
	A	●②③④⑤	①②③④	①②③	①②	①
	C	①②③④⑤	①②③④	①②③	①②	①
7. Work on Written Assignments	I	①②③④⑤	①②③④	①②③	●②	①

The nature of the data extracted from the CS is basically of two types. First the simple frequency of occurrence of any given activity, director, group type, or combinations of these factors can be computed for each snapshot, summed across snapshots and converted to a percentage based upon the total frequency of all events. These are reasonable indicators for characterizing the classroom setting, but fall short of accounting for how many students are actually involved in each configuration. The second type of information, therefore, weights the frequency of occurrence data by the estimated number of students involved using an algorithm based upon the known class size and the definitions of group sizes. Since the reliability results are similar for both types of scoring systems, only the results for the weighted (by students) frequencies will be presented here.

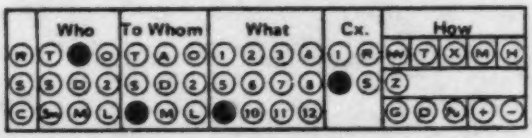
The five minute interaction portion of the observation record is a more continuous accounting of how time is spent in the classroom, focusing upon the teacher and the interactive process between teacher and students. Each interaction was recorded in the following FMI "frame," and an average of 60 such frames were recorded by the observers in a given five minute observation period:



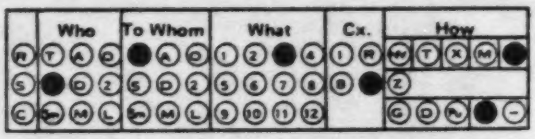
In effect, one of these optically scannable frames would be "bubbled-in" on the average of every 5 seconds depicting who was doing what to whom and how and in what context. For example, if the teacher (who) was correcting (what) a student (whom) with guidence (how) during instruction (context), the frame would be bubbled in by the observer as follows.



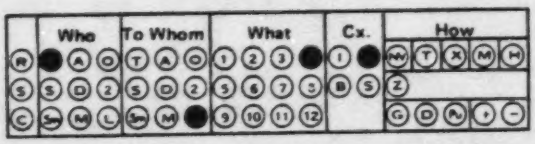
An aide correcting several students in the behavioral context (i.e., discipline and control) would be coded as follows:



A student responding to the teacher in a non-task and humorous "social" context with noticeable positive affect would be coded as follows:



The teacher explaining "routine" procedures to the total class would be coded as follows:



The nature of the data extracted from the FMI is basically of one type: for each "who-to whom-what-context-how" interaction defined, the

percentage of the total FMI frames compiled over the observation conforming to the interaction specifications is computed.

Clearly, the combinations and number of quantifiable pieces of information in the FMI and CS sections of the observation system are almost endless. For the purposes of this study, only marginal variables will be utilized, i.e., variables which are defined by holding all factors constant except for the one in question. For example, averages for each CS activity can be computed over all director and grouping types. Alternatively, they can be computed for each director over all activities and grouping types... and so on. Likewise, averages can be computed for each of the "who" codes in the FMI over all the remaining code categories...and so on. The Appendix to this report identifies each variable utilized in this study with a descriptive phrase and a code number for table reference.

Sample, Procedures, and Analytic Method

Included in the data base accumulated during A Study of Schooling are observational data on 132 elementary and 887 secondary classes. Classes were randomly sampled at each grade level at each elementary school; at secondary schools, classes were randomly sampled to represent each of eight major subject areas. The thirty-eight schools (13 elementary, 12 junior high or middle, and 13 senior high) represent a purposive sample of schools across the nation, with considerable variation in such factors as school size (pupil enrollment), economic status of the community, race/ethnicity of student body, and geographic location and characteristics.

Each class was usually observed by one assigned observer on three different occasions (full days at the elementary level and full periods at the secondary level) during a two-week period in the field. These data were accumulated over occasions into a single observational protocol for

each classroom for further research purposes. However, in 58 and 299 of the elementary and secondary classes respectively, a second observer was paired with the assigned observer for one of the three observation days (or periods), and these data were extracted for purposes of analyzing inter-observer reliability.

One easily computed coefficient of reliability is the Pearson product-moment correlation coefficient (r) between the pairs of assigned observers' ratings and those of the one-time-only "visiting" observers. This is a defensible analytical procedure since (a) the assigned observer was more familiar with the class in about two-thirds of the pairs and (b) the visiting observer was often the least well-trained of the pair. However, it can be readily argued that since all observers achieved training reliability criteria, they should, for all practical intents and purposes, be logically indistinguishable from one another. The intraclass correlation coefficient, then, becomes the more appropriate reliability coefficient. This index has the added advantage of being compatible with generalizability theory so long as observers can be thought of as a random sample from an infinitely large observer population. In this case, the intraclass correlation (r_{11}) is the reliability of the average single observer. The reliability of the average of k observers (r_{kk}) is simply r_{11} "blown-up" by a factor of k utilizing the Spearman-Brown prophesy formula. (r_{kk} would also be the alpha coefficient of internal consistency for k observers.) The tables to follow will present r, r_{11}, r_{22} and the predicted value of k such that $r_{kk} = .80$.[2]

Results and Discussion

Table 1 presents the correlational results between the paired observers for each variable defined in the Appendix. Before discussing

- - - - - - - - - - - - - -

Insert Table 1 about here

- - - - - - - - - - - - - -

them in detail, several points should be noted. First, the r and r_{11} values are nearly equal in most cases, indicating that the two raters are, for practical intents and purposes, interchangeable.

Second, the r_{11}, r_{22} and k values all represent the same inter-observer reliability information, but in different ways. If, for example, practical considerations dictated that no more than 3 observers could be used on any given occasion, than it would be necessary that $r_{11} > .57$ (or $r_{22} > .73$) assuming a desired reliability of at least .80 for the average observer rating. For the purposes of this report, inter-observer reliabilities resulting in k = 1, 2 and 3 will be described as highly, moderately and minimally reliable respectively; results for k > 3 will be described as low or unreliable.

Third, a number of variables have extremely high or extremely low (occasionally, even negative) inter-observer reliabilities because of restricted variances due to low frequencies (often near zero) of occurrence. For example, the inter-observer reliability for SNP48 at the secondary level is a perfect 1.0. However, the activity ("simulation, role playing") has an average relative frequency of occurrence of about one half of one percent. In effect, in 296 classes, observer pairs "agreed" on zero frequency of occurrence, while in the remaining 3 classes, the observer pairs happened to agree on what little non-zero results occurred. But if these few remaining pairs disagree, the correlation between them becomes negligible. Secondary results for variables SNP56 and FMI46 illustrate this effect. To facilitate interpretation, therefore, the means for each

variable are also included in the Appendix; and the reliability results for those CS and FMI variables with means close to zero percent should be interpreted accordingly. That is, the "reliability" of these variables may be more a function of their unobservable nature than of any consistency between the ratings of observers.

Finally, caution must be further excercised in interpreting results within and between variable subcategories of the CS and FMI due to the redundancies implicit in the instrument and variable definitions. Within subcategories (CS activities; FMI adult to student "whats"; etc.), the scores (i.e., percentages) are ipsative since they must always total 100%. Between subcategories, the same "micro" events may have been counted for the more "macro" variables defined. For example, FMI4 is the total percentage of all instructional frames and, therefore, includes all I-context frames entering into the definitions of subsequent FMI variables. The variable FMI15 almost always occurs in an instructional context representing, on the average, about 50% of all I-frames. The reliability data for variables FMI4 and FMI15, therefore, are clearly not independent.

Notwithstanding these interpretive difficulties, however, the following results reveal general trends in the dependability of the variables with practical implications for gathering observations at desired levels of inter-observer reliability.

Physical Environment Inventory

Only the half dozen most often utilized PEI variables were indluded for analysis here. The student-space ratio (PEI 1) was calculated by dividing the number of students enrolled by a classroom equivalence rating made by the observer. (One classroom equivalent was defined as the typical, self-contained classroom housing 30-35 students.) Apparently this was a

highly reliable observation at both elementary and secondary levels. Twelve different categories of equipment and materials (teaching machines, books and magazines, etc.) were available to check if they were observed in the classroom. The variable PEI2 is the sum of these checks and is highly reliable at the secondary level and moderately so at the elementary level. The four remaining variables (PEI3 to PEI6), pertaining to the degree of openness, flexibility in seating, presence of learning centers and alteration of physical environment are moderately to highly reliable. The only exception is variable PEI6 at the elementary level. Observers apparently had some difficulty judging the degree to which the basic elementary classroom was altered by novel bulletin board displays, plants, rugs, curtains, murals, etc.

Daily Summary

This is the most "subjective" part of the observation system as revised for the Study. Observers were required to make ratings based upon their impressions of more losely defined constructs such as "adequacy" of materials, "predominant" locus of decision-making, etc. Nevertheless, the general level of inter-observer reliability for these items is not that much different for those in other sections of the instrument. For some information, e.g., locus of decision-making items (DS4 - DS12), the coefficients are moderate to high; for others, e.g., space and adequacy and use of materials ratings (DS1 - DS3) the results are very poor.

Of particular interest is the summary index for locus of decision-making (DS11) which, in effect, is the average percentage of "predominately student" observations over seven loci. This measure is highly reliable at the elementary level and minimally so at the secondary level. This difference is most likely due to the extended length of time elementary

classes were observed (1 day) as compared to that (1 period) for secondary classes. Not being able to observe the locus of decision-making items (DS12) is minimally reliable at both schooling levels.

Classroom Snapshot

Generally, the results for both elementary and secondary levels are very similar, particularly if only variables with ample frequency of occurrence (say over 5%) are considered. For example, one of the more unreliable variables, the activity labeled "demonstration" (SNP46), had a near zero frequency of occurrence. Other "non-occurring" activities, like "simulation/role playing" (SNP48) or "adult disciplining students" (SNP56) either are perfectly reliable or totally unreliable for the statistical reasons noted above. In effect, the most interesting information here is how often these activities do not occur as agreed upon by both observers.

The activities involving the greater proportions of students--"preparation..." (SNP44), "explain..." (SNP45) and "work..." (SNP50)--have moderate to high reliability coefficients. Another classroom configuration, not directly involving students, is the teacher monitoring and/or observing students (SNP33). This was also a more frequently occurring variable with moderate inter-observer reliability. The several remaining non-student involved configurations (SNP32, SNP34 and SNP35) were less reliable and less frequent.

Recognizing what director-type(s) and what group-type(s) (SNP61-SNP67) is (are) involved in classroom activities appears to be a moderately to highly reliable observational task. The only exception is the minimally reliable result for the group type "one student" (SNP64). This is not

unexpected considering the greater difficulty of differentiating individuals (vs groups) in the classroom situation at any moment in time.

Finally, a number of other useful variables derived from the CS are shown to be moderately to highly reliable. Particularly surprising is the reliability of the rather subjective rating of "class interest level" (SNP15). The variety of activities and grouping variables (SNP16-SNP19) are also reasonably reliable. These variables are created simply by adding up the number of different activities in progress (by snapshot or by observation) and averaging (over snapshots or observations). Lastly, for elementary classes only, observers were required to code the subject area in progress when appropriate for activities occurring. This task appears to be highly reliable for the five basic subject areas observed (SNP68-SNP72).

Five Minute Interaction

Again, the results for these variables are not surprising considering the statistical implications for variables with low frequencies of occurrence. In the FMI section, however, less frequently occurring variables are usually a result of more specific variable definitions, i.e., more restrictions on the permissible codes for the who-to whom-what-context-how facets of the FMI frame.

Among the more "macro" FMI variables included in this analysis are the percentages of frames recorded in the instructional, behavioral and remaining (routine plus social) contexts. These variables (FMI4, FMI5 and FMI6) can be interpreted as estimates of the relative expenditures of classroom "time" in each context. These variables are moderately to highly reliable at the secondary level and minimally to moderately reliable at the elementary level. Note that FMI5 is one of the few low frequency macro

variables; nevertheless, when discipline and control situations do occur in the classroom, observers appear to agree reasonably well on their recording of "B"-contexts.

Other reliable macro variables include those estimating total verbal interaction and general affective climate. Adult initiated verbal interaction (FMI7) is highly reliable at both elementary and secondary levels; that for students (FMI32) is moderately and minimally so at the secondary and elementary levels respectively. The general affect variables (FMI51-FMI53) show similar reliability patterns, with one exception: general positive affect is quite unreliable at both elementary and secondary levels. Both general positive and negative affect variables have low frequencies of occurrence. However, as borne out by these data, the ability to train for and observe instances of negative affect is easier than the same for positive affect.

The remaining variables included for analysis here involve the specification of two or more FMI facets. Variables FMI8 through FMI19 include most of the possible adult (usually teacher) to student(s) "whats." The most frequently occurring "whats" are instruction/explanation (FMI15), monitoring/observing (FMI19), direct questioning (FMI8), and requests/commands (FMI11). These teacher behaviors exhibit moderate inter-observer reliabilities at both schooling levels, with the exception of "requests/commands" at the elementary level. Of the remaining low frequency variables only FMI16 evidences minimal reliability at the elementary level. This is particulary disappointing since "9" codes (correction) were emphasized in training for various research purposes.

The counterparts of these variables, student(s) to adult "whats" (FMI33 through FMI39), are generally less reliable. The most frequently

occurring "what" is student(s) responding (to teacher initiated question or request). Yet this variable (FMI34) is unreliable at the elementary level, although moderately so at the secondary level. The only other variables in this series with minimal to moderate inter-observer reliabilities are student(s) initiated interaction (FMI33) and student(s) questioning (FMI39), both under 5% in relative frequency of occurrence.

All the variables constructed to describe the nature of teacher-student(s) interactions, i.e., the "hows", are low frequency events. The only "hows" for teachers with reasonably consistent reliability at both schooling levels are those indicating movement (FMI23), use of materials (FMI24), and use of humor (FMI25). Interestingly, this latter "how" is extremely infrequent and is included in the general positive affect variable which is not very reliable. Apparently, among the infrequent contributors to positive classroom affect, humor may be the easiest to recognize consistently. None of the "hows" for students are particularly reliable and many have near zero occurrence rates.

Finally, a couple of other FMI codes are worth considering. An estimated ten percent of the "time" is spent by teachers working by themselves in non-instructional contexts. This variable (FMI20) is highly reliable at the secondary level and moderately so at the elementary level. Showing the same pattern of inter-rater reliabilities is FMI2, the percent of "repeat" frames. Although not of particular substantive interest when aggregated over observation protocols, it is useful to know that observers are using the "R" code at the beginning of each FMI frame in a generally consistent fashion.

Conclusions

Considering the complexity of the observational system, the inter-observer reliability results reported here are quite encouraging. With relatively few exceptions, variables defining behaviors which occur at a rate of at least five to ten percent or more have reasonable reliabilities, i.e., no more than 3 (and often only 1 or 2) observers are required to reach a reliability level of .80 for the average of the observers' ratings. Unreliability occurs mostly for those variables defining behaviors which are rarely observed. It must be reiterated, however, that observing the lack of occurrence of these variables is a consistent skill among observers. The unreliability stems from failure of observers to agree on the few times the behaviors do occur. To be sure, the unreliable results also signal failure during training of adequately defining codes and/or teaching observers to recognize the intended teacher-student behaviors. The implications of these data, then, are quite straightforward for further use of the observation system. For example, were we to use it again, we would, among other things, pay particular attention to the definition and use of "9" (correction) "what" codes and "+" and "-" (affect) "how" codes.

Finally, the infrequently occurring variables may simply represent observable facts regarding the constructs being measured. Perhaps to the reader's dismay, discussions regarding the substantive implications of the relative absence and presence of significant pedagogical processes has been avoided in this paper. For example, in this author's view, the relatively flat, neutral classroom affect, the preponderance of teacher talk, directing, instructing, explaining, and lecturing, and students working on written assignments and responding compliantly, all point to very uninspired

classroom learning environments which may adversely affect the quality of learning. However, much more needs to be discussed in these respects and that discussion is not the focus of this report. Moreover, the interpretation of the observational findings from A Study of Schooling is best done with the entire sample and full three-day observational protocols. Such studies are currently underway and will be the subject of future reports.

FOOTNOTES

1. More information on the background and current activities of A Study of Schooling can be found in the series of four articles published in the Phi Delta Kappan beginning with Goodlad, Sirotnik and Overman (1979).

2. Since correlations are mainly sensitive to the rank order of ratings of observer pairs, the mean differences between observer pair ratings are also useful indicators of agreement. Confidence intervals for these means should be small and include the point estimate of zero. An even more conservative test of observer agreement uses the absolute differences between observers' ratings. These analyses have been completed and strongly support the results and interpretations in this report. Tables of observer difference data are available from the author upon request.

REFERENCES

Baker, J.P., Giesen, P.A. and Norwood, C.R. Confusability: A criterion-referenced reliability procedure for interaction analysis observation instruments. Paper presented at the annual meeting of the American Educational Research Association, 1974.

Cronbach, L.J., Gleser, G.C., Nanda, H. and Rajaratnum, N. The dependability of behavioral measurements. New York: John Wiley and Sons, 1972.

Cronbach, L.J., Rajaratnum, N. and Gleser, G.C. Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 1963, 16, 137-163.

Giesen, P. and Sirotnik, K.A. The methodology of classroom observation in a study of schooling. A Study of Schooling, Technical Report No.5, 1979.

Goodlad, J.I., Sirotnik, K.A., and Overman, B.C. A study of schooling: An overview. Phi Delta Kappan, 1979, 61, 176-178.

Lindquist, E.F. Design and analysis of experiments in psychology and education. Boston: Houghton-Mifflin, 1953.

Medley, D.M. and Mitzel, H.E. Measuring classroom behavior by systematic observation. In Gage, N.L. (Ed.) Handbook of research on teaching. Chicago: Rand McNally, 1963.

Shavelson, R. and Demsey-Atwood, N. Generalizability of measures of teaching behavior. Review of Educational Research, 1976, 46, 553-611.

Stallings, J. and Giesen, P.A. A study of confusability of codes in observational measurement. Paper presented at the annual meeting of the American Educational Research Association, 1974.

Stallings, J. and Kaskowitz, D. Follow through classroom observation evaluation 1972-1973. SRI Project URU-7370, 1974, Stanford Research Institute.

TABLE 1

Inter-Observer Correlational Data for
Elementary and Secondary Classes

	Elementary				Secondary			
Variable	r	r_{11}	r_{22}	Estimate of k for r_{kk}=.80	r	r_{11}	r_{22}	Estimate of k for r_{kk}=.80
Physical Environment Inventory								
PEI1	1.00	1.00	1.00	1	.93	.92	.96	1
PEI2	.75	.74	.85	2	.83	.83	.91	1
PEI3	.88	.87	.93	1	.88	.89	.94	1
PEI4	.81	.80	.89	1	.83	.82	.90	1
PEI5	.76	.75	.86	2	.77	.75	.86	2
PEI6	.49	.48	.65	5	.76	.75	.86	2
Daily Summary								
DS1	.80	.79	.88	1	.65	.64	.78	3
DS2	1.00	1.00	1.00	1	.44	.44	.61	5
DS3	.39	.39	.56	7	.46	.45	.62	5
DS4	1.00	1.00	1.00	1	.62	.63	.77	3
DS5	.94	.94	.97	1	.78	.77	.87	2
DS6	*	*	*	*	.68	.68	.81	2
DS7	.81	.79	.88	2	.61	.60	.75	3
DS8	.54	.54	.70	4	.58	.57	.73	3
DS9	.70	.67	.80	2	.74	.74	.85	2
DS10	1.00	1.00	1.00	1	.72	.71	.83	2
DS11	.82	.82	.90	1	.66	.65	.79	3
DS12	.63	.61	.76	3	.66	.65	.79	3
Classroom Snapshot								
Type of Activity:								
SNP44	.92	.92	.96	1	.77	.77	.87	2
SNP45	.81	.80	.89	1	.87	.87	.93	1
SNP46	.45	.37	.54	7	.58	.56	.72	4
SNP47	.74	.65	.79	3	.79	.79	.88	2

TABLE 1 (Continued)

Variable	Elementary r	Elementary r_{11}	Elementary r_{22}	Elementary Estimate of k for r_{kk}=.80	Secondary r	Secondary r_{11}	Secondary r_{22}	Secondary Estimate of k for r_{kk}=.80
Classroom Snapshot								
Type of Activity:								
SNP48	.64	.57	.73	3	1.00	1.00	1.00	1
SNP49	.91	.90	.95	1	.94	.94	.97	1
SNP50	.81	.82	.90	1	.87	.87	.93	1
SNP51	.91	.90	.95	1	.83	.82	.90	1
SNP52	.75	.69	.82	2	.74	.72	.84	2
SNP53	.80	.75	.86	2	.85	.85	.92	1
SNP54	.95	.94	.97	1	.99	1.00	1.00	1
SNP55	.81	.79	.88	2	.82	.82	.90	1
SNP56	1.00	1.00	1.00	1	.02	.01	.01	>30
SNP60	.88	.87	.93	1	.72	.72	.84	2
Type of Director:								
SNP61	.88	.87	.93	1	.86	.85	.92	1
SNP62	.93	.92	.96	1	.76	.75	.86	2
SNP63	.85	.85	.92	1	.86	.85	.92	1
Type of Group:								
SNP64	.60	.59	.74	3	.66	.65	.79	3
SNP65	.93	.92	.96	1	.85	.85	.92	1
SNP66	.89	.89	.94	1	.69	.68	.81	2
SNP67	.91	.90	.95	1	.85	.85	.92	1
Misc. Other Codes:								
SNP15	.87	.85	.92	1	.67	.67	.80	2
SNP16	.94	.94	.97	1	.83	.83	.91	1
SNP17	.83	.83	.91	1	.73	.72	.84	2
SNP18	.89	.89	.94	1	.78	.79	.88	2
SNP19	.71	.71	.83	2	.72	.72	.84	2
SNP32	.88	.87	.93	1	.37	.35	.52	8
SNP33	.79	.79	.88	2	.69	.69	.82	2
SNP34	.56	.54	.70	4	.72	.71	.83	2
SNP35	.74	.72	.84	2	.50	.49	.66	5

TABLE 1 (Continued)

Variable	Elementary r	Elementary r_{11}	Elementary r_{22}	Elementary Estimate of k for r_{kk}=.80	Secondary r	Secondary r_{11}	Secondary r_{22}	Secondary Estimate of k for r_{kk}=.80
Content I.D. Summary:								
SNP68	.87	.87	.93	1	N/A	N/A	N/A	N/A
SNP69	.92	.92	.96	1	"	"	"	"
SNP70	.95	.94	.97	1	"	"	"	"
SNP71	.97	.96	.98	1	"	"	"	"
SNP72	.95	.96	.98	1	"	"	"	"
Five Minute Interaction								
Context Totals:								
FMI4	.60	.60	.75	3	.84	.83	.91	1
FMI5	.77	.77	.87	2	.80	.79	.88	2
FMI6	.63	.63	.77	3	.84	.83	.91	1
Adult to Student(s) "Whats":								
FMI8	.70	.69	.82	2	.75	.75	.86	2
FMI9	.44	.37	.54	7	.29	.29	.45	10
FMI10	.63	.63	.77	3	.45	.45	.62	5
FMI11	.53	.53	.69	4	.74	.74	.85	2
FMI12	.46	.45	.62	5	.42	.42	.59	6
FMI13	.51	.50	.67	4	.46	.45	.62	5
FMI14	.56	.54	.70	4	.20	.18	.31	18
FMI15	.68	.68	.81	2	.79	.79	.88	2
FMI16	.56	.57	.72	3	.53	.52	.68	4
FMI17	.67	.56	.72	4	.22	.22	.36	14
FMI18	.22	.19	.32	17	.00	.00	.00	>30
FMI19	.88	.87	.93	1	.76	.75	.86	2
Adult to Student(s) "Hows":								
FMI21	.34	.33	.50	8	.54	.53	.69	4
FMI22	.46	.30	.46	10	.17	.13	.23	27
FMI23	.62	.61	.76	3	.70	.69	.82	2
FMI24	.86	.87	.93	1	.80	.80	.89	1

TABLE 1 (Continued)

Variable	Elementary r	Elementary r_{11}	Elementary r_{22}	Elementary Estimate of k for r_{kk}=.80	Secondary r	Secondary r_{11}	Secondary r_{22}	Secondary Estimate of k for r_{kk}=.80
Adult to Student(s) "Hows":								
FMI25	.70	.69	.82	2	.72	.71	.83	2
FMI26	.40	.35	.52	8	.77	.77	.87	2
FMI27	.43	.42	.59	6	.36	.35	.52	8
FMI28	.60	.54	.70	4	.29	.29	.45	10
FMI29	.37	.36	.53	7	.98	.92	.96	1
FMI30	.55	.54	.70	4	.33	.32	.49	9
FMI31	.63	.59	.74	3	.32	.24	.39	13
Student(s) to Adult "Whats":								
FMI33	.59	.59	.74	3	.68	.68	.81	2
FMI34	.49	.48	.65	5	.73	.72	.84	2
FMI35	.39	.32	.49	9	.11	.09	.17	>30
FMI36	.47	.42	.59	6	.26	.25	.40	12
FMI37	.23	.22	.36	14	.36	.36	.53	7
FMI38	.09	.09	.17	>30	.40	.40	.57	6
FMI39	.62	.61	.76	3	.74	.74	.85	2
Student(s) to Adult "Hows":								
FMI40	.28	.28	.44	11	.61	.61	.76	3
FMI41	-.04	*	*	*	.41	.36	.53	7
FMI42	.35	.33	.50	8	.25	.24	.39	13
FMI43	.92	.90	.95	1	.55	.54	.70	4
FMI44	.09	.05	.10	>30	.29	.22	.36	14
FMI45	.40	.36	.53	7	.50	.50	.67	4
FMI46	.30	.28	.44	11	-.01	*	*	*
FMI47	*	*	*	*	*	*	*	*
FMI48	*	*	*	*	*	*	*	*
FMI49	.12	.07	.13	>30	.31	.13	.23	27
FMI50	.21	.20	.33	16	.58	.57	.73	3

TABLE 1 (Continued)

Variable	Elementary r	Elementary r_{11}	Elementary r_{22}	Elementary Estimate of k for r_{kk}=.80	Secondary r	Secondary r_{11}	Secondary r_{22}	Secondary Estimate of k for r_{kk}=.80
Total Verbal Interaction:								
FMI7	.85	.83	.91	1	.90	.90	.95	1
FMI32	.64	.64	.78	3	.73	.72	.84	2
Combined Affect:								
FMI51	.45	.39	.56	7	.53	.53	.69	4
FMI52	.64	.63	.77	3	.87	.83	.91	1
FMI53	.82	.80	.89	1	.82	.83	.91	1
Misc. Other Codes:								
FMI20	.74	.71	.83	2	.88	.89	.94	1
FMI2	.72	.72	.84	2	.81	.80	.89	1

* Insufficient variance for computations to proceed.

APPENDIX

VARIABLE DEFINITIONS

Variable Code	Description	Overall Mean Elementary	Overall Mean Secondary
	Physical Environment Inventory		
PEI1	Student-Space Ratio	27.66	23.77
PEI2	Equipment and Materials Available	6.72	3.18
PEI3	Presence of Adjacent Usable Space (1=no; 2=yes)	1.34	1.35
PEI4	Flexibility of Furnishings (1=flexible; 2=combination; 3=rigid)	2.29	2.04
PEI5	Learning Centers (1=none; 2=1 or 2; 3=3 or more)	1.44	1.09
PEI6	Degree of Alteration to Physical Environment (1=little or none; 2=moderate; 3=high)	1.85	1.39
	Daily Summary		
DS1	Space Rating (on a scale of 1=crowded, 2=adequate, 3=spacious)	1.82	2.12
DS2	Adequacy of Materials Rating (on a scale of 1=inadequate, 2=adequate)	1.98	1.98
DS3	Student Use of Materials & Equipment Rating (on a scale of 1=restricted, 2=partially restricted, 3=unrestricted)	1.63	1.94
	Locus of Decision-Making (on a scale of 1=predominantly teacher, 2=predominantly student) Regarding:		
DS4	Seats	1.04	1.33
DS5	Groups	1.05	1.41
DS6	Content	1.00	1.03
DS7	Materials	1.06	1.09
DS8	Space	1.08	1.11
DS9	Time	1.03	1.06
DS10	Learning Activities	1.02	1.04
DS11	Average Locus of Decision-Making Over all 7 Areas	1.04	1.13
DS12	Average Percentage of Unobservable Loci Over all 7 Areas	7.66	25.39

VARIABLE DEFINITIONS (Continued)

Variable Code	Description	Overall Mean Elementary	Overall Mean Secondary
	Classroom Snapshot		
	Type of Activity:		
SNP44	Preparation for assignments or instructions/cleanup	15.62	16.18
SNP45	Explain, lecture, or read aloud	22.63	25.55
SNP46	Demonstration	0.78	1.54
SNP47	Discussion	5.30	7.02
SNP48	Simulation, role playing	0.05	0.44
SNP49	Reading	6.69	3.17
SNP50	Work on written assignments	26.06	16.37
SNP51	Practice or perform (psycho-motor, physical)	6.88	10.19
SNP52	Practice or perform (verbal)	4.10	4.17
SNP53	Taking test or quiz	2.02	4.73
SNP54	Audio visual equipment	5.38	2.43
SNP55	Student non-task behavior or teacher social interaction with student(s)--no assignment	4.53	7.32
SNP56	Adult disciplining students	0.25	0.03
SNP60	Student non-task behavior during subject	2.93	1.93
	Type of Director:		
SNP61	Teacher/aide (combined)	56.45	57.10
SNP62	Student/cooperative (combined)	8.81	8.99
SNP63	Independent	34.89	33.00
	Type of Group:		
SNP64	One student	1.35	3.96
SNP65	Small group of students	7.12	9.13
SNP66	Medium/large group of students	25.80	15.82
SNP67	Total class	65.91	70.17

VARIABLE DEFINITIONS (Continued)

Variable Code	Description	Overall Mean Elementary	Overall Mean Secondary
	Miscellaneous Other Codes:		
SNP15	Percentage of class at high interest level (Mean rating on a scale of 1=0-24%, 2=25-49%, 3=50-74%, 4=75-100%)	3.35	3.40
	Variety of Instructional Activities:		
SNP16	By snapshot	1.10	0.92
SNP17	By observation	4.24	1.76
	Variety of Grouping:		
SNP18	By snapshot	1.36	1.26
SNP19	By observation	3.82	1.91
	Adult:		
SNP32	Disciplining students	0.32	0.12
SNP33	Monitoring and/or observing students	10.16	8.56
SNP34	Custodial, routines	6.35	7.11
SNP35	Not involved or out of classroom	0.29	0.46
	Content Identification Summary for Elementary Only: (Sum of Occurrences Across Activities):		
SNP68	English	12.14	N/A
SNP69	Mathematics	4.02	N/A
SNP70	Science	0.72	N/A
SNP71	Social Studies	1.69	N/A
SNP72	The Arts	1.97	N/A
	Five Minute Interaction		
	Context Totals (Anyone to Anyone):		
FMI4	Instruction	72.72	76.79
FMI5	Behavior	4.96	1.66
FMI6	Routine and social (combined)	22.30	21.53

VARIABLE DEFINITIONS (Continued)

Variable Code	Description	Overall Mean Elementary	Overall Mean Secondary
	Adult to Student(s) "Whats":		
FMI8	Direct Questions	7.58	6.30
FMI9	Open Questions	0.49	0.60
FMI10	Response	2.83	3.81
FMI11	Imperative (requests/commands)	8.60	5.22
FMI12	Comments/general action	3.39	4.30
FMI13	Acknowledgment/praise (specific)	3.30	2.19
FMI14	Encouragement/praise (general)	1.03	0.66
FMI15	Instruction/explanation	19.10	29.93
FMI16	Correction (with/without guidance)	5.98	2.90
FMI17	No response	0.06	0.08
FMI18	Refusal (flat/with reason)/rejection	0.06	0.04
FMI19	Monitor/observe	8.34	7.79
	Adult to Student(s) "Hows":		
FMI21	Nonverbal	5.72	6.21
FMI22	Touching	0.27	0.17
FMI23	With movement	4.19	5.81
FMI24	Materials used	2.88	8.06
FMI25	With humor	0.10	0.39
FMI26	Life experience examples	0.22	0.51
FMI27	With guidance	1.67	1.25
FMI28	Demeaning/threatening	0.16	0.04
FMI29	Punishing	0.07	0.07
FMI30	Positive affect	0.82	0.45
FMI31	Negative affect	0.24	0.11
	Student(s) "Whats" (Adult Involved):		
FMI33	Directive (student initiated interaction)	3.86	4.97
FMI34	Compliant (response)	13.43	9.44
FMI35	Refusal	0.03	0.03
FMI36	Not responding	0.61	0.51
FMI37	Contributing	3.34	2.89
FMI38	Assertive	0.41	0.39
FMI39	Questioning	3.10	4.42

VARIABLE DEFINITIONS (Continued)

Variable Code	Description	Overall Mean Elementary	Overall Mean Secondary
	Student(s) to Adult "Hows":		
FMI40	Nonverbal	3.36	2.31
FMI41	Touching	0.01	0.01
FMI42	With movement	1.56	0.81
FMI43	Materials used	0.77	1.82
FMI44	With humor	0.04	0.13
FMI45	Life experience examples	0.18	0.07
FMI46	With guidance	0.01	0.01
FMI47	Demeaning	0.00	0.00
FMI48	Punishing	0.00	0.00
FMI49	Positive Affect	0.11	0.17
FMI50	Negative Affect	0.02	0.08
	Total Verbal Interaction:		
FMI7	Adult initiated (to one or more students)	50.13	53.33
FMI32	Student initiated (with and without adult responding)	24.18	22.52
	Combined Affect:		
FMI51	General positive	1.33	1.26
FMI52	General negative	0.55	0.46
FMI53	General neutral	80.69	80.32
	Miscellaneous Other Codes:		
FMI20	Adult working alone	10.17	11.02
FMI2	Percent of valid frames "repeated" (type of interaction unchanged)	33.22	43.31